DOCUMENT RESUME

ED 107 702 TM 004 527

AUTHOR Roudabush, Glenn E.
TITLE Models for a Beginning Theory of Criterion-Referenced Tests.
PUB DATE [Apr 74]
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Criterion Referenced Tests; Decision Making; *Evaluation Methods; Measurement Techniques; *Models; *Objective Tests; Scores; Statistical Analysis; Student Evaluation; Test Construction; *Testing; Test Reliability; Test Validity; True Scores

ABSTRACT

In this paper, several models for the psychometric nature of criterion-referenced tests are presented and results derived with implications for test construction, reliability and validity measures, and educational decision making. Both dichotomous and continuous underlying abilities to perform are considered. Illustrative data fitting both cases is also presented. (Author)

ED107702

MODELS FOR A BEGINNING THEORY OF CRITERION-REFERENCED TESTS

by

Glenn E. Roudabush

CTB/McGraw-Hill

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A paper presented at the

National Council for Measurement in Education meetings in Chicago

April 1974

TM 004 527

MODELS FOR A BEGINNING THEORY OF CRITERION-REFERENCED TESTS

A frequently quoted definition of a criterion-referenced test is this one given by Glaser and Nitko (1971): "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." I endorse this definition, particularly in the context of this paper. They go on to say: "Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain, and such measurements are referenced directly to this domain for each individual measured." Defining the class or domain of tasks is the function of an instructional objective. A well written, or "good," instructional objective will give a reasonably unambiguous definition of the domain it is intended to specify, so that one will be able to tell whether or not a given item or task is within the domain or outside of it. Note that the requirement is that "measurements are taken on representative samples of tasks drawn from . . . (the) domain." The samples are not (necessarily) randomly drawn. For many significant instructional objectives, the domain of tasks cannot be ennumerated and, indeed, may be infinite in number. The wording of the instructional objective, though, should be clear enough so that it is possible to construct tasks or test items that represent the domain with some degree of confidence.

Some writers (e.g., Emrick, 1971; Livingston, 1972) use the phrase "criterion-referenced test" to indicate a collection of items all of which are intended to measure the same objective. The definition above allows the phrase to indicate a collection of items that measure an organized set of

related objectives giving as many scores as objectives represented in the test. When the latter is the case, important collateral information may be available in the test itself to improve the accuracy of the scores for the represented objectives (see, for example, Roudabush and Green, 1972; Humbleton and Novick, 1973). Either interpretation may be appropriate depending upon the intended use of the score(s) or the nature of the subject under discussion.

An item on a paper and pencil test, when completed by a student, is a sample of that student's behavior. Furthermore, the item is a sample selected from some universe of all possible behaviors which might have been selected to represent or measure some particular domain of behaviors. The more limited the domain, the easier it is to select behaviors to represent it, and more confidence can be placed in the representativeness of the selected behaviors for the domain. If we assume a Platonic truth about a student with respect to his or her ability to perform the behaviors described by an objective and also assume a level of specificity of the objective such that (ideally) if a student can perform correctly one behavior from the domain, then he or she can perform them all, then for any given item on a test there are four possible outcomes: (1) the student cannot perform the behavior and does not get the item correct; (2) the student cannot perform the behavior but does get the item correct -- a false positive score; (3) the student can perform the behavior but does not get the item correct -- a false negative score; and (4) the student can perform the behavior and does get the item correct. If a large number of students were repeatedly tested with items sampling the domain of behaviors specified by such an objective, the proportion of correct responses of individual students should approach a stable bimodal distribution. The lower mode would give the mean proportion for

students who cannot perform the behavior and the upper mode would give the mean proportion for students who can perform the behavior.

Implicit in this kind of specific objective is the idea that the student will be able to give a perfect performance every time or else he has not mastered the objective. In the process of determining mastery, some pragmatic standard of performance will have to be imposed and a margin for errors in classification tolerated. These are practical problems of measurement, however, (for this kind of specific all-or-none objective) and not inherent in the objective as such. We really do want the student to be able to correctly add whole numbers all of the time and not just 90% of the time, but that is not to say that he or she will always do so.

Most writers on criterion-referenced testing concern themselves with the concept of mastery or non-mastery of objectives, but implicitly or explicitly assume some underlying continuum of performance within the domain of an objective. The criterion of mastery on this continuum becomes of major concern. Humbleton and Novick (1973), for example, are concerned to more accurately estimate a student's standing with respect to a cut-point. Hively, Maxwell, Rabehl, Sension, and Lundin (1973) and Millman (1972) are concerned to estimate the number of items in a domain which a student "really" can correctly answer. Millman quotes Ebel (1971) as follows: "... abilities, understandings, and appreciations are in the experience of almost everyone, not all-or-none adaptations. They are matters of degree. None but the simplest of them can ever be mastered completely by anyone [p. 287]" and professes agreement with that position. Undoubtedly many abilities, understandings, and appreciations are a matter of degree, but I believe there are many that are not a matter of degree. They are, in fact, all-or-none occurrences and they are not necessarily simple or unimportant. Whether a measure of an objective seems continuous or dichotomous

probably depends upon the specificity of the objective and, in part, on the nature of the content of the objective. A measure of an objective specifying a heterogenous domain may seem to reflect a continuum of ability, but, in fact, is made up of many dichotomous sub-objectives. Tabulating scores across individuals will give a distribution of scores that looks continuous. Further, Ebel is denying sudden insight, the "ah-haa" experience that is, I would hope, in the experience of almost everyone.

I am proposing, then, two models of the underlying nature of what is measured by a criterion-referenced test, each of which applies in some cases and not in others. The first assumes an underlying all-or-none, dichotomous, "true" score and the second assumes an underlying continuous "true" score. The word "true" in "true" score was placed in quotation marks because I want to differentiate it from the usual interpretation it is given. For example, in mathematics it will generally not be satisfactory to know that a student at a particular time in a particular situation did, in fact, correctly add two 3-digit numbers four times out of five. What we really wish to know is whether or not he or she is able to do the addition consistently over a long period of time with accuracy, that is, we wish to infer something about the state of the examinee with respect to his newly aquired ability to do addition. We are still concerned with potentially observable behavior and not with internal traits, dispositions, or values. Our concern is with a potential or ability to behave in particular ways, which is one step removed from direct observation.

Now consider Figure 1 which I have labeled Case 1: a dichotomous measure of a dichotomous true score. The probability of making an error, where an error is defined as classifying a person as having mastered the objective

when, in fact, he has not or as classifying a person as having not mastered the objective when, in fact, he has, is shown in the figure as the shaded portions and is equal to $P(X = 1 \mid T = 0) + P(X = 0 \mid T = 1)$. If the dichotomous measure is a test item, then this figure represents the theoretical item characteristic curve.

It is interesting to note that Klein and Cleary (1967) have shown that for a dichotomous measure of a dichotomous true score (which they call a Platonic true score) when both measure and true score take values of zero or one and error for an individual is therefore -1, 0, or +1, that the true score and error are negatively correlated and that the error can have a mean of zero only when the number of false positives equal the number of false negatives in the sample (which is unlikely). They also show that a second dichotomous measure of the same true score will have errors correlated with the errors in the first measure. This, of course, violates some key assumptions of classical test theory and, among other things, causes inflated estimates of reliability using classical methods. Werts, Linn, and Jöreskog (1973), using congeneric test theory (Jöreskog, 1971), have shown that if the error score is allowed to take values other than -1, 0, and +1, then the classical assumptions can be met. Doing so, however, begins to make the true score look less than Platonic.

Figure 2, which I have labeled Case 2: a pseudo continuous measure of a dichotomous true score, shows the case where some number, N, of measures of the true score are used to obtain an observed score, X, for an individual. In this case, there is a distribution of observed scores for people with a true score of zero and another distribution of observed scores for people with a true score of one. The end of the solid lines are meant to indicate the mean or mode of the two distributions. In this case, in order to get an estimate of the true score, a criterion observed score, X_c, or cut-point

must be established as indicated. The probability of an error of classification is then: $P(X \geq X_c \mid T = 0) + P(X < X_c \mid T = 1)$. The error is shown as the shaded area in the figure.

If, in ignorance of the true score, we simply plot the frequency distribution of the observed score, then we combine the two distributions of Figure 2 into one and would expect (with the hypothesis of a dichotomous true score) that the distribution would be bimodal or, perhaps, U-shaped. Such a distribution is shown as the solid line in Figure 3. The 20 item criterion test in this figure was constructed to measure a single objective from the Prescriptive Mathematics Inventory (Gessel, 1972). The additional lines in Figure 3 will be described in a moment.

Suppose that we have two independent measures of the same objective. Call one a CRT, scored zero or one, and the other a criterion, also scored zero or one. The criterion may be direct observation, teacher rating, or another test. We can then form the following table of observed frequencies:

		Criterion 0	Criterion 1	
CRT	0	f_{00}	f_{01}	$f_{0\cdot}$
	1	f_{10}	f_{11}	$f_{1\cdot}$
		$f_{\cdot 0}$	$f_{\cdot 1}$	N

Where f_{00} is the number of cases not showing mastery of the objective on either the CRT or the criterion, f_{01} is the number of cases not showing mastery on the CRT, but showing mastery on the criterion, and so on. N is

the total number of cases in the sample. For a dichotomous true score, there is some true number of cases, say N_0, who have, if fact, not mastered the objective and some other number of cases, say N_1, who have mastered the objective. The theoretical table of frequencies is then:

Criterion

CRT	0	1	
0	N_0	0	N_0
1	0	N_1	N_1
	N_0	N_1	N

Now let $\alpha_1 = P(X \geq X_c \mid T = 0)$ = the probability that non-masters show mastery on the CRT,

$\alpha_2 = P(X \geq X_c \mid T = 0)$ = the probability that non-masters show mastery on the criterion,

$\beta_1 = P(X < X_c \mid T = 1)$ = the probability that masters show non-mastery on the CRT, and

$\beta_2 = P(X < X_c \mid T = 1)$ = the probability that masters show non-mastery on the criterion.

From these definitions and the joint frequency tables, it can be shown that:

$$\left.\begin{aligned} f_{00} &= N_0(1 - \alpha_1)(1 - \alpha_2) + N_1\beta_1\beta_2 \quad , \\ f_{01} &= N_0(1 - \alpha_1)\alpha_2 + N_1\beta_1(1 - \beta_2) \quad , \\ f_{10} &= N_0\alpha_1(1 - \alpha_2) + N_1(1 - \beta_1)\beta_2 \quad , \text{ and} \\ f_{11} &= N_0\alpha_1\alpha_2 + N_1(1 - \beta_1)(1 - \beta_2) \quad . \end{aligned}\right\} \qquad [1]$$

Only three of these equations are independent, since any one of the frequencies can be obtained by subtracting the sum of the remaining three from N, the fixed sample size. The system is under-determined since we have only three equations to solve for the five unknowns: α_1, α_2, β_1, β_2, and N_0 or N_1 (since $N = N_0 + N_1$).

If we assume that $\alpha_2 = \beta_2 = 0$, that is that the criterion admits of no error, then the following solutions obtain since there are now but three unknowns:

$$\left.\begin{aligned} \alpha_1 &= \frac{f_{10}}{f_{00} + f_{10}} \quad , \\ \beta_1 &= \frac{f_{01}}{f_{11} + f_{01}} \quad , \\ N_0 &= f_{00} + f_{10} \quad , \text{ and} \\ N_1 &= N - N_0 = f_{11} + f_{01} \quad . \end{aligned}\right\} \qquad [2]$$

Consider the following table of observed frequencies:

		Criterion 0	Criterion 1	
CRT	0	143	39	182
	1	45	291	336
		188	330	518

In this table, the criterion score is the 20 item test whose distribution appears in Figure 3 dichotomized at $X_c = 11$ and the CRT score is a single item from the Prescriptive Mathematics Inventory (PMI) that corresponds to the same objective. The values of α_1, β_1, N_0, and N_1 from equations 2 are:

$$\alpha_1 = .24,$$
$$\beta_1 = .11,$$
$$N_0 = 188, \text{ and}$$
$$N_1 = 330.$$

In this case, the probability of making a false positive classification is about two times the probability of making a false negative classification. In Figure 3, the dashed line gives the distribution of the 20 item criterion test for those students who correctly answered the corresponding PMI item and the line with hash marks gives the distribution for those students who incorrectly answered the corresponding PMI item. Taking the criterion test as error free, the combined probability of misclassifying a student on the basis of the single PMI item is .35 and the probability of correctly classifying a student is $1 - \alpha_1 - \beta_1 = .65$. The latter may be taken as an index of reliability for the one item test. Given the situation in which these data were collected, it is likely that the number of false positive classifications is inflated. Over a two week testing period in which the PMI was always administered first, it is likely that fatigue effects account for some of the false positives. It is also possible that the number of false negative classifications is inflated due to learning on the part of some students. This last effect is accentuated in other distributions from the same study.

Figure 4 shows the distribution of a 15 item criterion test with the distributions for those students who showed mastery on a five item CRT and those students who showed non-mastery on the five item CRT. Mastery was defined as four out of five items correct. These distributions came from data collected for the PIRAMID project (Project: Individualized Reading And Mathematics, Inter-District). This project was initiated by a consortium of California school districts. This objective is specific and has the U-shaped form indicative of an underlying dichotomous true score. In this case, taking the criterion test as error free, there is one false negative classification based on the five item CRT. The probability of misclassification is .01 and the index of reliability is .99. Needless to say, such results are quite uncommon.

It should be noted that if three independent measures of the same objective are available, then all relevant parameters for the three variables can be computed. There are seven parameters: α_1, α_2, α_3, β_1, β_2, β_3, and N_0 and seven independent equations available from the two by two by two cube of observed frequencies.

Figure 5, which I have labeled Case 3: a dichotomous measure of a continuous true score, shows a traditional kind of item characteristic curve (if the dichotomous measure is an item scored 0 or 1) with the addition of an assumed criterion true score, θ, indicating mastery of the objective. The probability of making an error of classification in this case is indicated by the shaded area and is equal to $P(X = 1 \mid T < \theta) + P(X = 0 \mid T \geq \theta)$. Empirical item characteristic curves, where scores on a pool of items written to measure one objective are substituted for true score, could be useful in item selection when a reasonable cut score, X_c, has been established. In this case, characteristic curves which cross the cut score near their center should be more sensitive to instruction than curves which do not. The

importance of sensitivity to instruction in criterion-referenced test item selection has been discussed elsewhere (Cox and Vargas, 1966; Roudabush, 1973).

Figure 6, labeled Case 4: a pseudo continuous measure of a continuous true score, is the situation assumed and most discussed in the literature on criterion-referenced testing. It requires that a criterion observed score, X_c, be established and that a criterion true score, θ, be assummed. The probability of making an error of classification is shown by the shaded area and is equal to: $P(X \geq X_c \mid T < \theta) + P(X < X_c \mid T \geq \theta)$. The solid line in Figure 7 shows the distribution of a nine item objective score which would seem to fit this model. Also plotted is the distribution for students who showed mastery and non-mastery on a three item measure of the same objective where the criterion of mastery was set at two out of the three items correct. The items for both scores were taken from the tryout data for the Prescriptive Reading Inventory (1972). Notice that, if the nine item test is taken as an error free criterion test, there is no cut score that does not result in large errors of classification. For this reason, in the published PRI we trichotomize objective scores leaving a middle ground for "review" between mastery and non-mastery of each objective in the test.

In this last case, a step function approach to estimating error given any particular value of θ such as that of Reed (undated) or Humbleton and Novick (1973) is appropriate.

References

Cox, R. C. and Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the National Council for Measurement in Education meetings in Chicago, February, 1966.

CTB Staff. Prescriptive Reading Inventory. Monterey, Calif.: CTB/McGraw-Hill, 1972.

Ebel, R. L. Criterion-referenced measurements: limitations. School Review, 1971, 79, 282-288.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Gessel, J. Prescriptive Mathematics Inventory. Monterey, Calif.: CTB/McGraw-Hill, 1972.

Glaser, R. and Nitko, A. J. Measurement in learning and instruction. In: Thorndike, R. L. (Ed.) Educational Measurement. Washington, D. C.: American Council on Education, 1971.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., and Lundin, S. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the Minnemast Project. CSE Monograph Series in Evaluation No. 1, Center for the Study of Evaluation, Univer. of Calif. at Los Angeles, 1973.

Humbleton, R. K. and Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, 109-134.

Klein, D. F. and Cleary, T. A. Platonic true scores and error in psychiatric rating scales. Psychological Bulletin, 1967, 68, 77-80.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Millman, J. Passing scores and test lengths for domain-referenced measures. Cornell Univer., 1972.

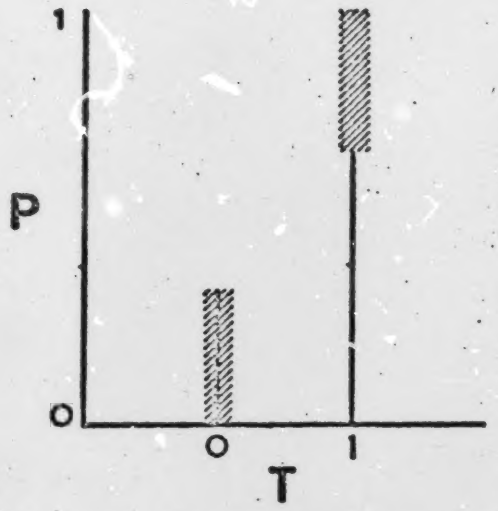
Reed, S. C. A step function test model for criterion and selection tests. Opinion Research Corp., undated paper.

Roudabush, G. E. Item selection for criterion-referenced tests. Paper presented at the American Educational Research Association meetings in New Orleans, February, 1973.

Roudabush, G. E. and Green, D. R. Aspects of a methodology for creating criterion-referenced tests. Paper presented at the National Council for Measurement in Education meetings in Chicago, April, 1972.

Werts, C. E., Linn, R. L., and Jöreskog, K. A congeneric model for Platonic true scores. Educational and Psychological Measurement, 1973, 33, 311-318.

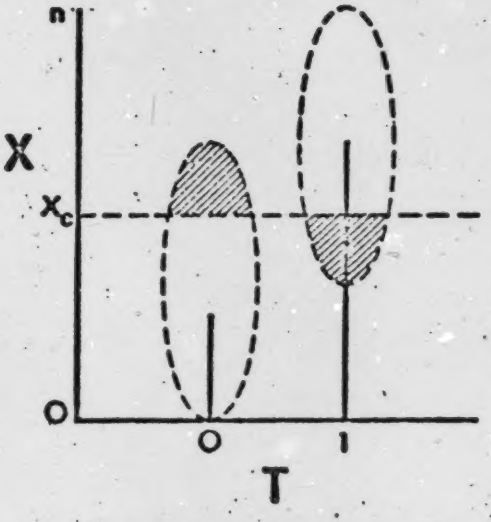
CASE 1: DICHOTOMOUS MEASURE – DICHOTOMOUS TRUE SCORE



$$P(E) = P(X = 1 \mid T = 0) + P(X = 0 \mid T = 1)$$

Figure 1. A dichotomous measure of a dichotomous true score showing the probability of making an error of classification based on the dichotomous measure.

CASE 2: PSEUDO CONTINUOUS MEASURE – DICHOTOMOUS TRUE SCORE



X_c = CRITERION OBSERVED SCORE

$$P(E) = P(X > X_c \mid T = 0) + P(X < X_c \mid T = 1)$$

Figure 2. A pseudo continuous measure of a dichotomous true score showing the probability of making an error of classification based on the pseudo continuous measure.

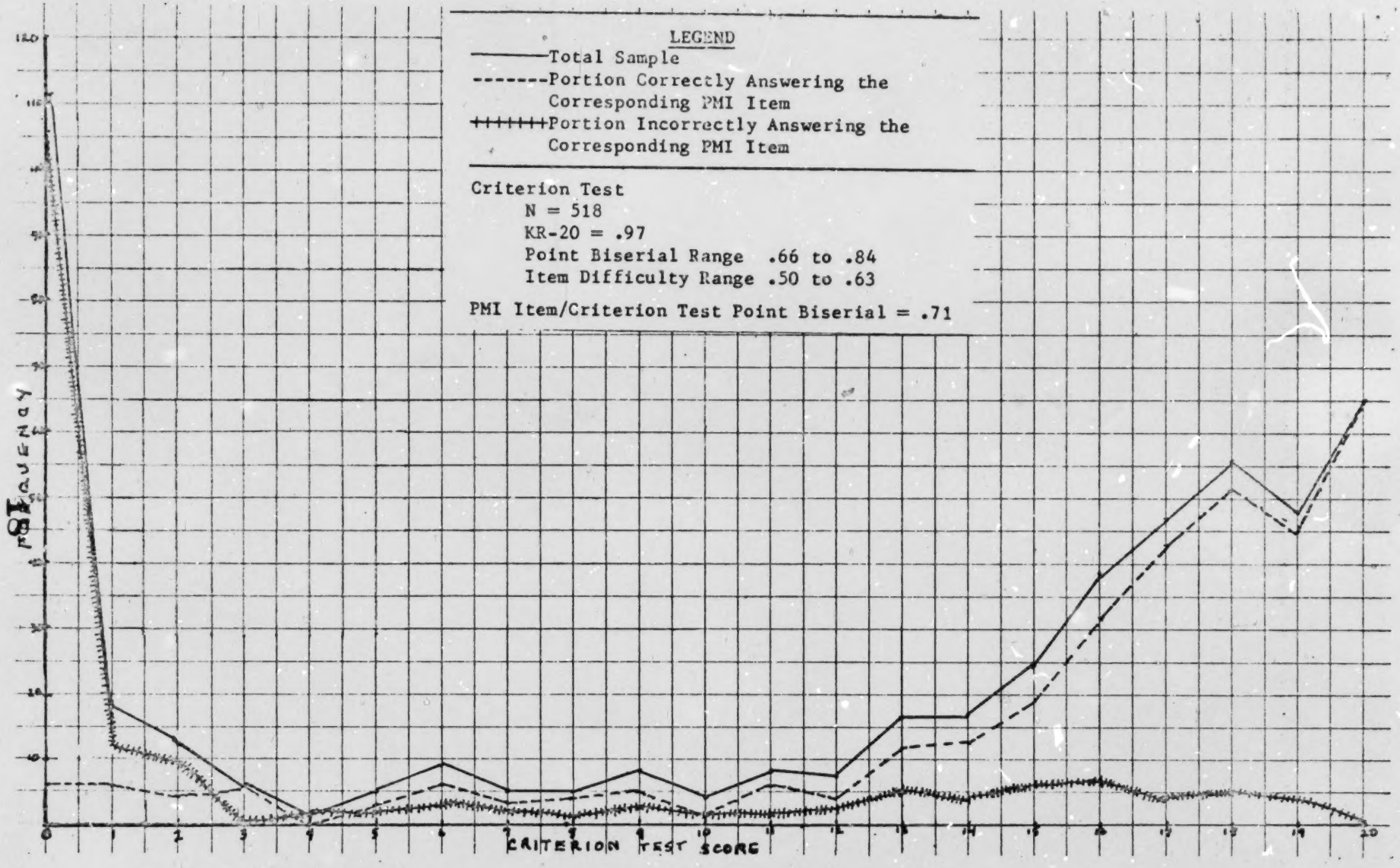

LEGEND
———Total Sample
-------Portion Correctly Answering the Corresponding PMI Item
+++++++Portion Incorrectly Answering the Corresponding PMI Item
Criterion Test
N = 518
KR-20 = .97
Point Biserial Range .66 to .84
Item Difficulty Range .50 to .63
PMI Item/Criterion Test Point Biserial = .71
FREQUENCY
CRITERION TEST SCORE

Figure 3. Distribution of Scores on PMI Criterion Test B-2: ADDITION OF THREE POSITIVE FRACTIONS for (1) The Total Sample, (2) That Portion of the Sample Correctly Answering the Corresponding PMI Test Item, and (3) That Portion of the Sample Incorrectly Answering the Corresponding PMI Item.

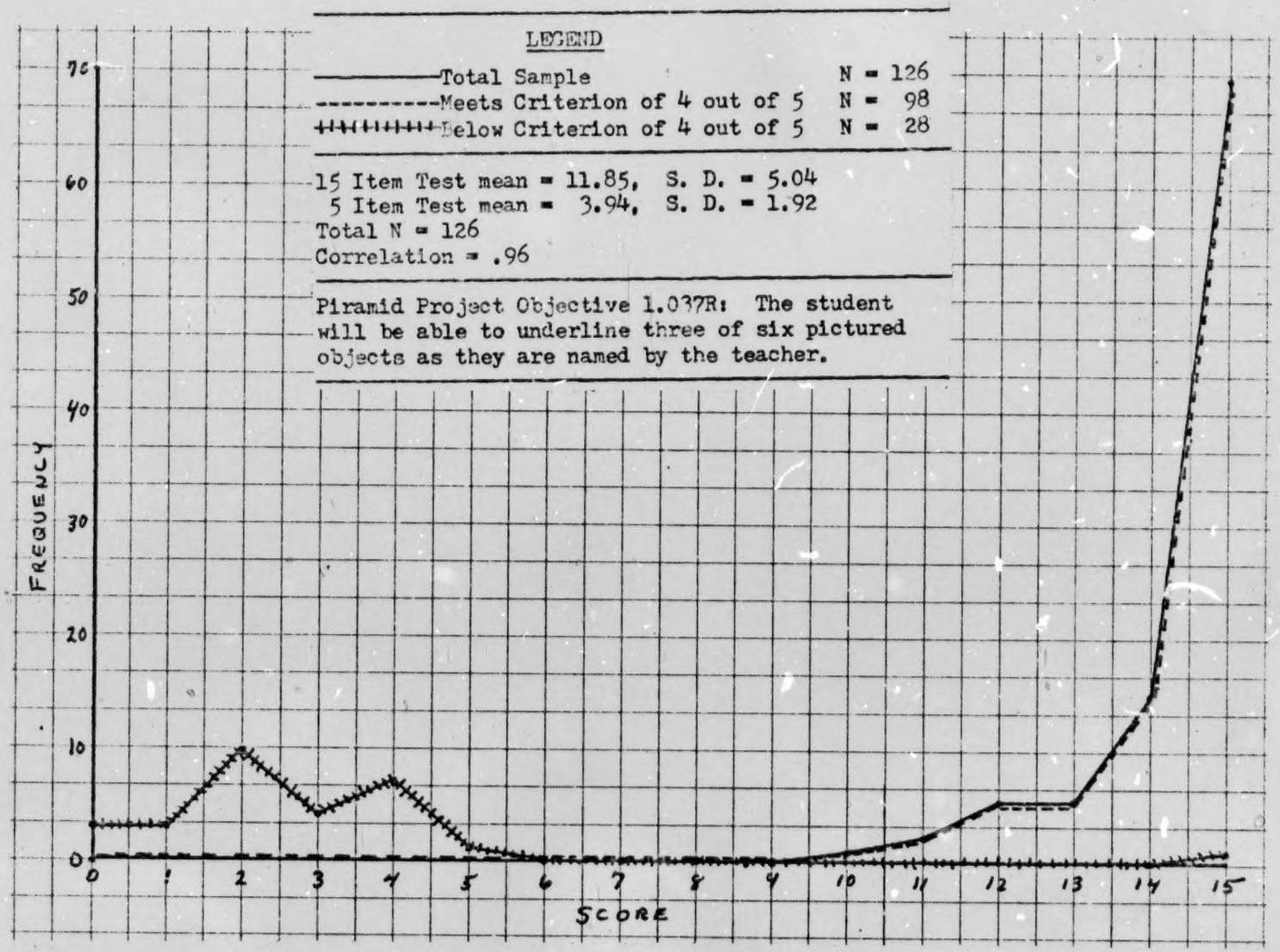
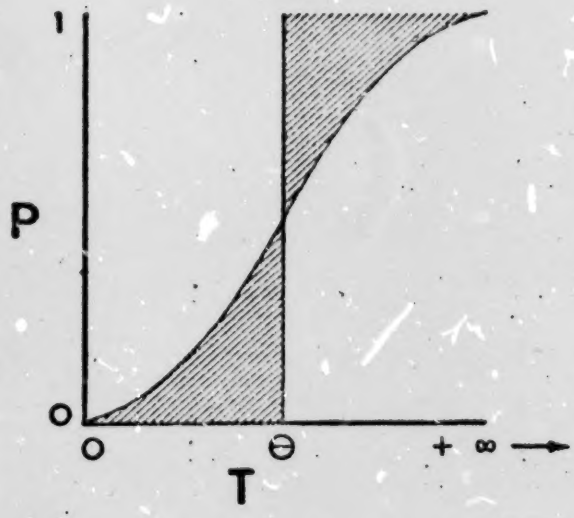

LEGEND
Total Sample N = 126
Meets Criterion of 4 out of 5 N = 98
Below Criterion of 4 out of 5 N = 28
15 Item Test mean = 11.85, S. D. = 5.04
5 Item Test mean = 3.94, S. D. = 1.92
Total N = 126
Correlation = .96
Piramid Project Objective 1.037R: The student will be able to underline three of six pictured objects as they are named by the teacher.
FREQUENCY
SCORE
0
10
20
30
40
50
60
70
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Figure 4. Distribution of scores on PIRAMID objective 1.037R: PICTURE VOCABULARY showing (1) the total sample, (2) that part of the sample not meeting the criterion of 4 out of 5 items correct, and (3) that part of the sample meeting the criterion. Used by permission of the PIRAMID Consortium, Paulla Hyatt, chairman.

CASE 3: DICHOTOMOUS MEASURE – CONTINUOUS TRUE SCORE

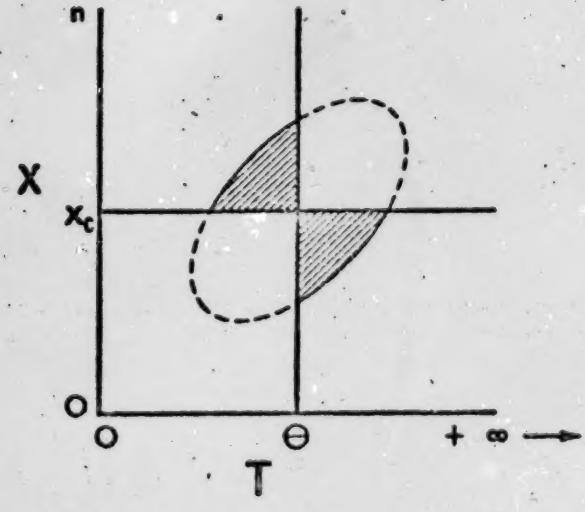


θ = CRITERION TRUE SCORE

$$P(E) = P(X = 1 \mid T < \theta) + P(X = 0 \mid T > \theta)$$

Figure 5. A dichotomous measure of a continuous true score showing the probability of making an error of classification based on the dichotomous measure given criterion true score θ.

CASE 4: PSEUDO CONTINUOUS MEASURE – CONTINUOUS TRUE SCORE



X_c = CRITERION OBSERVED SCORE

θ = CRITERION TRUE SCORE

$$P(E) = P(X > X_c \mid T < \theta) + P(X < X_c \mid T > \theta)$$

Figure 6. A pseudo continuous measure of a continuous true score showing the probability of making an error of classification based on the pseudo continuous measure with criterion X_c and given criterion true score θ.

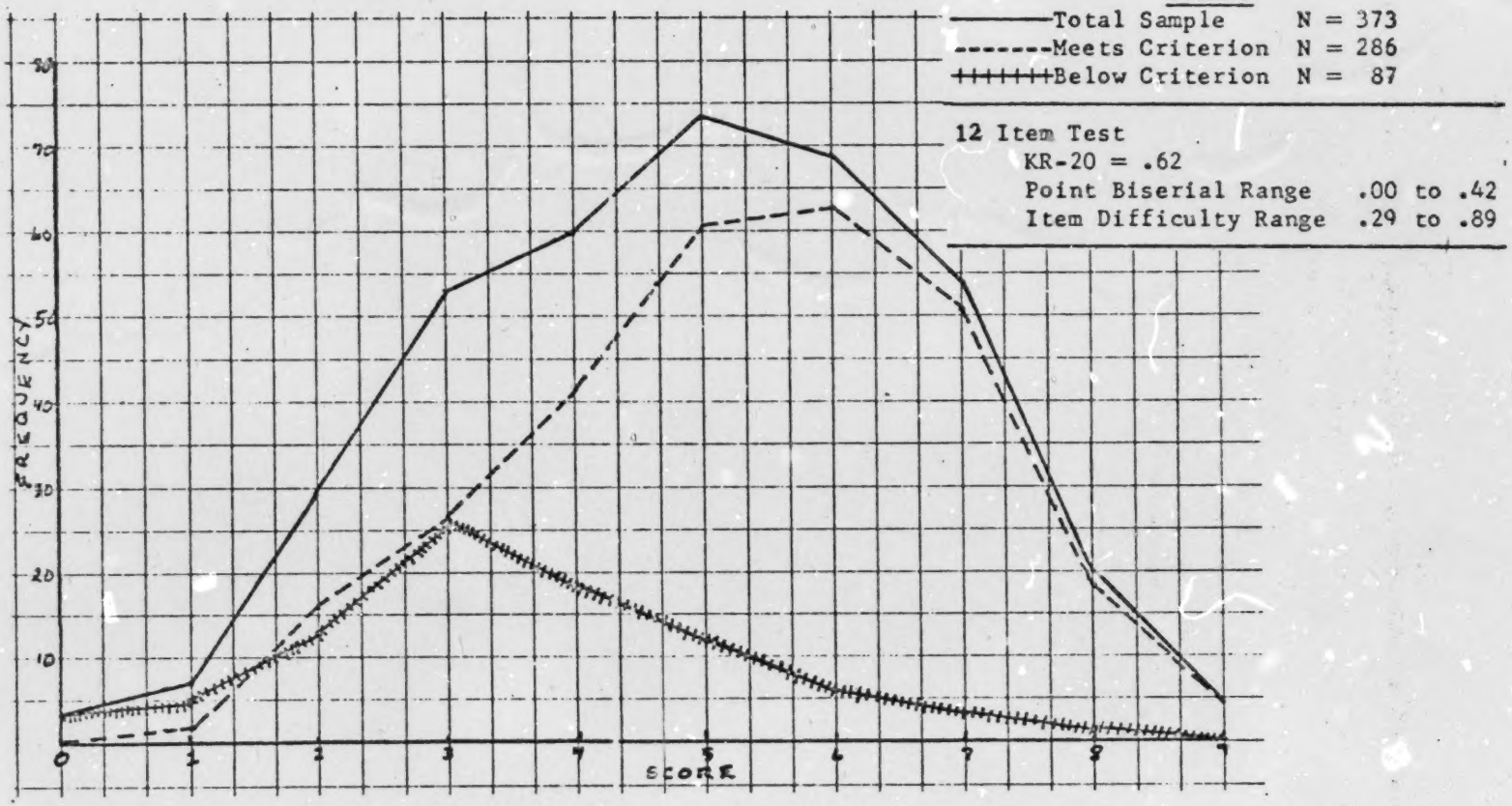


Figure 7. Distribution of Scores on PRI Objective 6, Level B: EVENT SEQUENCE Showing (1) The Total Sample, (2) That Part of the Sample Not Meeting the Criterion of 2 Out of 3 Items Correct, and (3) That Part of the Sample Meeting the Criterion. The Correlation Between the 3 Randomly Selected Items and the Remaining 9 Items is .46.